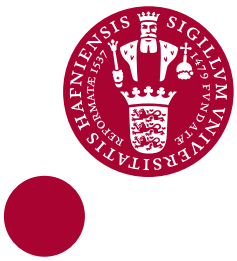

The Copenhagen Team Participation in the Check-Worthiness Task of the Competition of Automatic Identification and Verification of Claims in Political Debates of the CLEF-2018 CheckThat! Lab

Hansen, Casper; Hansen, Christian; Simonsen, Jakob Grue; Lioma, Christina

Published in:
CLEF 2018 Working Notes

Publication date:
2018

Document version
Publisher's PDF, also known as Version of record

The Copenhagen Team Participation in the Check-Worthiness Task of the Competition of Automatic Identification and Verification of Claims in Political Debates of the CLEF-2018 CheckThat! Lab

Casper Hansen, Christian Hansen, Jakob Grue Simonsen, and Christina Lioma

Department of Computer Science, University of Copenhagen (DIKU)
{c.hansen,chrh,simonsen,c.lioma}@di.ku.dk

Abstract. We predict which claim in a political debate should be prioritized for fact-checking. A particular challenge is, given a debate, how to produce a ranked list of its sentences based on their worthiness for fact checking. We develop a Recurrent Neural Network (RNN) model that learns a sentence embedding, which is then used to predict the check-worthiness of a sentence. Our sentence embedding encodes both semantic and syntactic dependencies using pretrained *word2vec* word embeddings as well as part-of-speech tagging and syntactic dependency parsing. This results in a multi-representation of each word, which we use as input to a RNN with GRU memory units; the output from each word is aggregated using attention, followed by a fully connected layer, from which the output is predicted using a sigmoid function. The overall performance of our techniques is successful, achieving the overall second best performing run (MAP: 0.1152) in the competition, as well as the highest overall performance (MAP: 0.1810) for our contrastive run with a 32% improvement over the second highest MAP score in the English language category. In our primary run we combined our sentence embedding with state of the art check-worthy features, whereas in the contrastive run we considered our sentence embedding alone.

Keywords: political debates, recurrent neural networks, sentence embedding, check-worthiness

1 Tasks Performed

The Copenhagen team participated in both Tasks 1 and 2 of the CLEF 2008 Fact Checking Lab for the English language. This report details our methods and results for *Task 1*, where we focused on the English task. Our participation in Task 2 is described in [5].

The aim of Task 1 is to identify sentences in a political debate that should be prioritized for fact-checking: given a debate, the goal is to produce a ranked list of all sentences based on their worthiness for fact checking.

One of the two examples given by the lab organizers [4] is shown in Table 1, where Hillary Clinton mentions Bill Clinton's work in the 1990s, followed by a claim made by Donald Trump stating that president Clinton approved the North American Free Trade Agreement (NAFTA). This last statement by Trump is worth checking and is flagged as such. We refer to the competition description [4] for further details on the competition and the provided dataset.

Table 1. Example of Check-Worthiness

Speaker	Sentence
CLINTON:	I think my husband did a pretty good job in the 1990s.
CLINTON:	I think a lot about what worked and how we can make it work again...
TRUMP:	Well, he approved NAFTA... (check-worthy)

2 Main Objectives of Experiments

The task of check-worthiness is different from the typical task of fact-checking because distilled claims are not provided, but rather the complete debate dialog is given. This means that most sentences will not be worthy of checking. Consequently, we reason that a rich representation of a sentence is needed in order to determine whether it is check-worthy.

The main objective of our experiments is to consider check-worthiness as two points that should *both* be satisfied:

1. The sentence content should be factually interesting
2. The sentence content should be possible to check

The first point requires that check-worthy sentences should contain semantic information that is interesting to verify as being true or not. For example, it is interesting to determine if Bill Clinton did approve NAFTA in the 1990s, but it would not be interesting to know what he had for breakfast. If a sentence is interesting, then later fact-checking is only possible if it is actually feasible to check the statement of the sentence, which is the requirement of the second point. Consequently when both points are satisfied, it follows that the sentence, or a part of it, should be syntactically structured in a way that resembles that of a check-able claim. This latter point, as well as the two original points, motivates our approach for learning a combined semantic and syntactical sentence embedding. Our approach to modelling this is described next.

3 Approaches Used and Progress Beyond State-of-the-Art

Our approach in this task is to learn a sentence embedding that can be used to predict the check-worthiness of a sentence. To do so, we encode the semantics

and the syntactic dependencies, both of which will be explained next. We encode the semantics of a sentence by the use of word embeddings, specifically *word2vec* [3], where each word is mapped into a vector space with the property that semantically similar words are close to each other. Furthermore, we employ Part-of-Speech (POS) tagging to better model the role of each word in the sentence. To encode the syntactic structure of a sentence we employ syntactic dependency parsing with the purpose of better encoding the syntactic structure a sentences needs to resemble in order to be checkable. Specifically, the dependency parsing maps each word to its dependency (as a tag) in relation to the sentence structure. Both POS tags and syntactic dependencies are represented using a one-hot-encoding.

For each word in a sentence, we thus have 3 distinct encodings that, together, represent the word. We use this *multi-representation* of each word as input to a recurrent neural network (See Figure 1) with GRU memory units, where the output from each word is aggregated using attention, followed by a fully connected layer, from which the output is predicted using a sigmoid function. GRU was chosen as opposed to the popular LSTM unit, since they have fewer parameters and therefore better suited to a small data setting as in this competition.

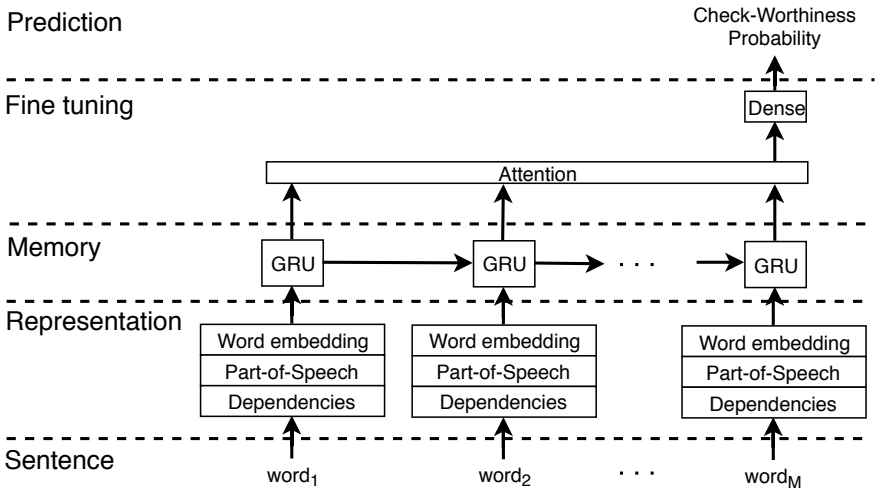

Fig. 1. Network architecture for our sentence embedding for a sentence with M words.

Related Check-Worthiness methods Existing state-of-the-art check-worthiness methods are based on feature engineering, by extracting a number of varied features from each sentence and potentially its context. So far actually learning (sentence) embeddings have not been done. ClaimBuster [2] focus on extracting features of sentiment, sentence length, TF-IDF, Part-of-Speech tags, and en-

tity extraction. The features are made on a sentence-level, and no information about the context of the sentence with regards to its surrounding sentences is used. Gencheva et al. [1] extend ClaimBuster by incorporating *context-aware* features into the representation, as well as more sentence-level features. Among other things, they extract contextual features such as the sentence position in a speaker segment, whether the speaker mentions the opponent, audience reactions, and a sentence's similarity to the segments around it. However, one of the features employed by Gencheva et al. is the largest overlap between the sentence, and sentences from *PolitiFact* and the training corpus. Due to the rules of the competition, the feature of the largest overlap from a sentence to *PolitiFact* claims is not included as a feature.

Compared to the above, our approach can be considered a sentence-level feature, since we do not include information from other sentences when generating the embedding. Thus, we also evaluate the combination of our sentence embedding with the context-aware baseline. Our choice of focusing the embedding entirely on the sentence and not the context was due to the low amount of training data.

4 Resources Employed

In our sentence embedding based approach we use the model described in Figure 1, where the GRU has 100 hidden units and the densely connected layer has 25 neurons and uses a ReLU activation function. For the word embedding we use a pretrained *word2vec* model based on Google News[1], and keep it fixed during training to avoid overfitting. For POS tagging and syntax dependency parsing we use spaCy[2], and for the neural network implementation Keras[3]. We apply no preprocessing to the raw sentences, except for tokenization, and do not remove stopwords. The reason for this is stopwords might contain important information for representing the syntactical structure of the sentence.

For ClaimBuster and the context-aware baseline we implement the features described in the original papers. Claimbuster provides an online API which we do not use due to it having worse performance compared to self-training[1]. For the classification algorithm we use the best observed model from [1], which is a feedforward neural network with 2 layers of 200 and 50 neurons respectively, both with ReLU activations. Following the original papers we apply stopword removal before generating the TF-IDF bag-of-words features, but generate all remaining features without stopwords removed. When evaluating the combination of our sentence embedding with the context-aware baseline, we combine the models by concatenating the second last layers, such that the final prediction is based on a 75 dimensional vector.

[1] `https://code.google.com/archive/p/word2vec/`
[2] `https://spacy.io/`
[3] `https://keras.io/`

5 Analysis of the Results

For evaluation of the models we use 3-fold evaluation where 2 of the 3 debates act as training data and the remaining as test. We follow the competition guidelines and report the MAP measure for each of the test debates individually and as an average. Additionally, we compute the MAP scores on each speaker by only considering the speakers sentences. This is done in order to determine if some of the models are better at adapting to certain individuals.

The performance for each model evaluated on the debates is shown in Table 2. ClaimBuster, based on sentence-level features, performs the worst. We see that by introducing context-aware features from the model proposed by Gencheva et al. [1] the MAP is improved in the first debate, but similar in the remaining. Our approach of training a combined semantic and syntactic sentence embedding performs better, and the lowest MAP score is increased from around 0.04 to 0.105 on the vice-presidential debate. However, it performs 0.09 worse on the first debate, and 0.09 better on the second. By combining our approach with the context-aware model we obtain the best model and obtain an average MAP of 0.158.

The performance of each model evaluated on the sentences by the individual speakers (without the moderators) is shown in Table 3. In this experiment we use the same train and test setup as before, but report the AUC for each speaker across the three runs. As before, the context-aware model is in most cases a direct improvement upon ClaimBuster. We observe that our sentence embedding performs notably better on the speakers from the vice presidential debate, and on Donald Trump, but worse on Hillary Clinton. By combining the embedding with the context-aware model the average MAP is slightly improved compared to only using the context-aware model, but using only our embedding overall obtains the best MAP score averaged over all speakers.

Table 2. MAP scores for each model across the provided training debates.

Model	1st-president	2nd-president	vice-president	Average
ClaimBuster [2]	0.159	0.109	0.043	0.104
Gencheva et al. [1]	0.193	0.114	0.039	0.115
Our	0.109	0.206	**0.105**	0.140
Our + Gencheva et al. [1]	**0.225**	**0.207**	0.043	**0.158**

5.1 Analysis

Across the debates we observe a large variation in performance depending on the used model: ClaimBuster obtains the lowest average MAP, but also uses the most simple features. The introduction of context-aware features provided a direct improvement by significantly improving the performance on the first

Table 3. MAP scores for each model across the speakers

Model	Trump	Clinton	Pence	Kaine	Average
ClaimBuster [2]	0.128	0.139	0.074	0.041	0.095
Gencheva et al. [1]	0.143	**0.201**	0.057	0.039	0.110
Our	0.225	0.077	**0.169**	**0.086**	**0.139**
Our + Gencheva et al. [1]	**0.245**	0.149	0.054	0.050	0.125

presidential debate. From these types of features we see an indication of limited generalizability, since they are not able to obtain good performance on the vice presidential debate, where the speakers are not present in the training debates. In our approach we do not directly encode specific features that might generalize poorly, but instead focus purely on deriving an embedding based on the semantics and syntactic dependencies of each sentence. The goal of this was to improve generalizability, which we also observe in the results, where we are able to increase the performance of the vice presidential debate by more than 0.06. When comparing our approach to the context-aware model we observe that they have orthogonal strengths, in the sense that our approach obtains higher scores when the context-aware model performs worse and vice versa. In the combination of the two approaches the performance on the two presidential debates increases and the highest MAP scores are obtained, however the performance on the vice presidential debate remains at the level of the context-aware model. We further consider the generalizability of the models by inspecting the performance on each of the individual speakers. It is interesting to consider that our approach perform noticeable better on the Trump/Pence duo compared to Clinton/Kaine, whereas the context-aware model and the combined approach favors Trump and Clinton compared to the vice presidential candidates. When considering the average MAP scores, then we similarly observe an increased generalizability of our method, since our sentence embedding alone obtains the best score averaged over all speakers.

5.2 Discussion of test data performance

Donald Trump was present in 6 out of the 7 provided testing debates and speeches, Hillary Clinton in 2, and a person not in the training data (Bernie Sanders) only in 1. For the speakerwise performances we believed the combined approach to be the optimal choice, since this obtained the best combined performance on both Trump and Clinton. As the contrastive run we submitted our proposed sentence embedding alone as it performed very similar to the combined approach and due to the previous arguments of improved generalizability, in the case that the Trump speeches are significantly different from the debates he participated in. The final results on the test data can be seen in Table 4. The combined approach ranked second in the task with a MAP of 0.1152, whereas our sentence embedding alone obtained the highest overall performance in the

task with a MAP of 0.1810. Compared to the second highest performing submission in the English language category our sentence embedding outperformed it by 32.5% (MAP: 0.1366). We believe the reason for the large performance difference is due to the improved generalizability of our sentence embedding, as the difference between the training debates compared to the debates *and* speeches in the testing data proved to be non negligible.

Our method does not use any context-based or global information, but is entirely reliant on the sentence that is being ranked. When inspecting the training data manually it is easy to see that an awareness of the overall context is necessary for ranking many of the sentences. Since our method which is working entirely on the sentences alone perform the best, it shows that the context-aware models may be prone to overfitting, and with the amount of training data available, it is better to focus entirely on what can be achieved from the sentence in isolation. The test data also contains multiple speeches in comparison to only debates in the training data. How the context should be treated in speeches and debates are very likely different, and therefore the models trained on context in debates does most likely not generalize well on all the test data.

Table 4. MAP scores for the submitted models on the testing data

Model	Test MAP
Our + Gencheva et al. [1]	0.1152
Our	0.1810

6 Perspectives for Future Work

In the future we plan to extend our model by incorporating context. At the moment our sentence embedding is based on each sentence alone, but as discussed previously, many check-worthy sentences are dependent on their context. One way to do this would be to create a context window embedding around the sentence in focus, such that the model is aware of what is said before and after. Inspired by Gencheva et al. [1], this could also be based on a person-specific context window embedding, by modelling what each debate participator has individually said. Naturally this leads to a more complex model, thus most likely requiring more training data than was provided in this competition.

References

1. P. Gencheva, P. Nakov, L. Màrquez, A. Barrón-Cedeño, and I. Koychev. A context-aware approach for detecting worth-checking claims in political debates. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 267–276. INCOMA Ltd., 2017.

2. N. Hassan, F. Arslan, C. Li, and M. Tremayne. Toward automated fact-checking: Detecting check-worthy factual claims by claimbuster. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1803–1812. ACM, 2017.
3. T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.
4. P. Nakov, A. Barrón-Cedeño, T. Elsayed, R. Suwaileh, L. Màrquez, W. Zaghouani, P. Gencheva, S. Kyuchukov, and G. Da San Martino. Overview of the CLEF-2018 lab on automatic identification and verification of claims in political debates. In *Working Notes of CLEF 2018 – Conference and Labs of the Evaluation Forum*, CLEF '18, Avignon, France, September 2018.
5. D. Wang, J. G. Simonsen, B. Larsen, and C. Lioma. The Copenhagen Team Participation in the Factuality Task of the Competition of Automatic Identification and Verification of Claims in Political Debates of the CLEF-2018 Fact Checking Lab. Technical report, CLEF Fact Checking Lab, 2018.